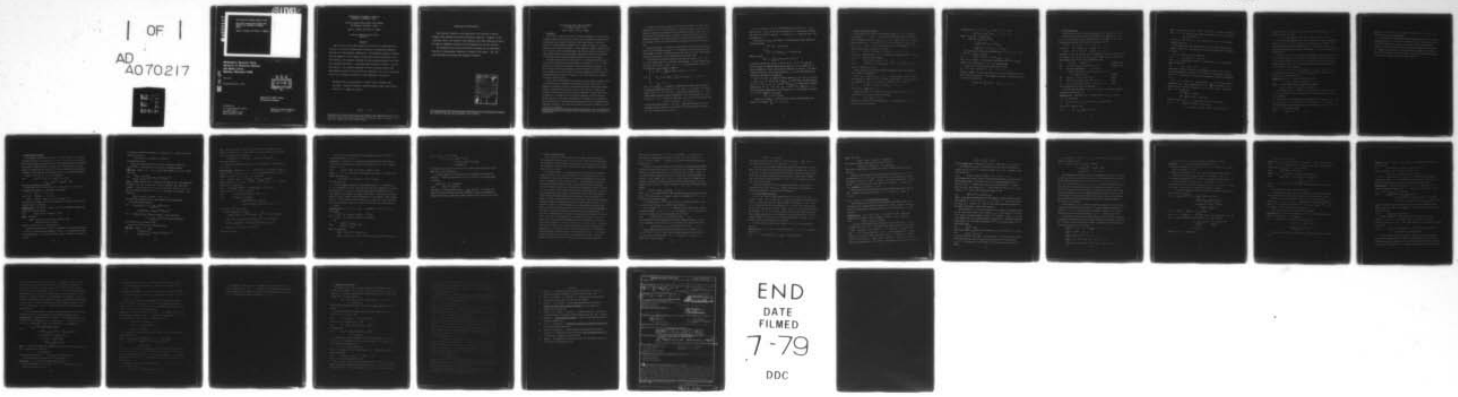
AD-A070 217 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
EFFICIENT HIGHER ORDER SINGLE STEP METHODS FOR PARABOLIC PROBLE--ETC(U)
MAY 79 J H BRAMBLE, P H SAMMON DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1958-PT-1 NL



(12) LEVEL II

MRC Technical Summary Report #1958

EFFICIENT HIGHER ORDER SINGLE STEP METHODS FOR PARABOLIC PROBLEMS: PART I

James H. Bramble and Peter H. Sammon

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

May 1979

Received March 15, 1979

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D. C. 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

EFFICIENT HIGHER ORDER SINGLE STEP METHODS FOR PARABOLIC PROBLEMS: PART I

James H. Bramble and Peter H. Sammon

Technical Summary Report # 1958
May 1979

ABSTRACT

Some efficient, high order methods are discussed for approximating the solution of an initial boundary value problem for a homogeneous parabolic equation with time dependent coefficients. The methods are based on Galerkin-type approximations in the spacial variables and single step methods in the time variable. The equations defining the time stepping procedure are solved only approximately however. A preconditioned iterative technique is used for this purpose. The resulting algorithm is shown to produce optimal order approximations using only the order of work required by the single step method applied to the parabolic problem with time independent coefficients.

AMS(MOS) Subject Classifications: Primary 65M15, Secondary 65N30

Key Words: Parabolic Equations, Galerkin Methods, Higher Order Methods

Work Unit #7 - Numerical Analysis

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. This material is based upon work supported by the National Science Foundation under Grants No. MCS76-07236 A02 and MCS78-09525.

Significance and Explanation

Many physical situations can be modelled by the solutions of initial boundary value problems for partial differential equations. Examples of such situations arise in the theory of heat conduction and other diffusion processes. The physical parameters involved are often dependent on the time variable.

The construction of accurate, efficient algorithms for the approximate solution of these parabolic equations is studied in this paper. New, efficient algorithms are proposed and completely analyzed.

EFFICIENT HIGHER ORDER SINGLE STEP METHODS FOR PARABOLIC PROBLEMS: PART I

James H. Bramble and Peter H. Sammon

I. Introduction

In this paper we study efficient ways to calculate approximations to the solution of a parabolic equation that are of third or fourth order in time and of high order in space. The approximations are generated by rational function based schemes (cf. Nassif and Descloux [7] or Baker, Bramble and Thomée [2] if the operator in question is time independent) but these schemes are modified in a manner suggested by Douglas, Dupont and Ewing [4] in their work on the Crank-Nicolson method. We study the schemes in the context of a linear parabolic equation with time dependent coefficients in this part of the work and we will generalize these schemes to nonlinear equations in Part II of this work.

The schemes suggested by Nassif and Descloux in [7] are single step methods based on a certain class of rational function approximations to the function e^{-x}, $x \in \mathbb{R}$ and a given family of discrete spacial operators. Nassif and Descloux give estimates in [7] that show that the resulting approximations make errors that are of optimal order. However the schemes are not really suitable for practical computation since each step of the time-stepping procedure involves the solution of a new linear system that is related to the family of spacial operators.

Douglas, Dupont and Ewing address this problem in [4] and suggest the remedy of using a preconditioned iterative technique to approximately solve the linear system. This approach only requires the solution of linear systems involving a fixed discrete spacial operator if sufficiently many iterations are done at each time step. Conditions are discussed in [4] that also allow one to iterate a fixed number of times at each time step (a number that is independent of the discretization parameters) and still observe the optimal order errors. The overall work required by this strategy is of the magnitude of the work required by the usual Crank-Nicolson scheme applied to a linear problem with time independent coefficients. Thus under these conditions, they obtain a scheme that is efficient as well as effective.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. This material is based upon work supported by the National Science Foundation under Grants No. MCS76-07236 A02 and MCS78-09525.

The results of this paper are similar to those of [4] with regard to our higher order schemes and are in fact in some ways stronger. In particular, because the schemes which we consider are inherently more dissipative than the Crank-Nicolson scheme, we are able to obtain the best results unconditionally. In addition, our analysis shows that the special closeness requirements of the initial values to the "elliptic projection" demanded in [4] are unnecessary and that a more natural and more easily computed initial projection may be used.

An alternate approach to the problem of finding efficient time stepping algorithms can be found in work by Douglas and Dupont in [3]. They analyze two efficient schemes for parabolic problems. The first is a method which is of first order in time (the Laplace-modified procedure) and the second is a three level method of second order in time.

We now introduce the parabolic problem and some convenient notation. Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a compact domain with a sufficiently smooth boundary $\partial\Omega$ and an outward pointing unit normal $\underline{n}(x) = (n_1(x),\ldots,n_d(x))$. Let $\tau > 0$. The following parabolic problem will be studied under certain smoothness assumptions:

$$(1.1)\quad \begin{cases} -u_t = L(t)u \equiv -\sum_{i,j=1}^{d} D_i(a_{ij}(x,t)D_j u) + a_o(x,t)u \quad \text{on} \quad \Omega \times (0,\tau), \\ u\big|_{\partial\Omega} = 0 \quad \underline{\text{or}} \quad \underline{n}A\nabla u\big|_{\partial\Omega} = \sum_{i,j=1}^{d} n_i\, a_{ij}\, D_j\, u\big|_{\partial\Omega} = 0 \quad \text{on} \quad \partial\Omega \times (0,\tau), \\ u\big|_{t=0} = v \quad \text{on} \quad \Omega. \end{cases}$$

Here $A = [a_{ij}(x,t)]_{i,j=1}^{d}$ is a symmetric, uniformly positive definite family of matrices of sufficiently smooth coefficients on $\bar{\Omega} \times [0,\tau]$, $a_o(x,t)$ is a nonnegative, sufficiently smooth function on $\bar{\Omega} \times [0,\tau]$ and $v(x)$ is a given initial data function on Ω. If the Neumann boundary conditions are under consideration, we will further require that $a_o(x,t)$ not vanish on $\bar{\Omega} \times [0,\tau]$ and that the coefficients $\{a_{ij}\}$ have the following special form:

$$a_{ij}(x,t) = a(x,t)\tilde{a}_{ij}(x), \quad 1 \leq i,j \leq d,$$

for sufficiently smooth functions a and $\{\tilde{a}_{ij}\}$. (This extra requirement ensures time independent boundary conditions). We will refer to the Dirichlet boundary conditions as BC_D and to the Neumann conditions as BC_N.

We let H^{ℓ} denote the usual $L^2(\Omega)$-based Sobolev space with norm $\|\cdot\|_{\ell}$, ℓ a non-negative integer. We also let H_o^1 denote the subspace of H^1 that consists of functions that vanish (in the sense of trace) on $\partial\Omega$. We will use $(\cdot,\cdot)$ to denote the usual $L^2(\Omega)$-inner product on Ω.

The operators $\{L(t)\}_{0\le t\le T}$ form a family of $L^2(\Omega)$-selfadjoint elliptic operators on the following domain:

$$D_L = \begin{cases} H^2 \cap H_o^1 & \text{if we have } BC_D . \\ H^2 \cap \{w \in H^2 : \underline{n}A\nabla w\big|_{\partial\Omega} = 0\} & \text{if we have } BC_N . \end{cases}$$

Moreover, the form

$$D(t)(\cdot,\cdot) = \sum_{i,j=1}^{d} (a_{ij} D_j(\cdot), D_i(\cdot)) + (a_o(\cdot),(\cdot))$$

is (strongly) coercive over $H_o^1 \times H_o^1$ if we have BC_D or $H^1 \times H^1$ if we have BC_N. Thus we can apply the standard parabolic equation theory to (1.1) (cf. Friedman [5] or Lions and Magenes [6]) and get the existence and uniqueness of solutions u for various classes of initial data. We will always assume that $v \in D_L$ and further smoothness and compatibility conditions will be added later.

We let $T(t): L^2(\Omega) \to D_L$ denote the solution operator for $L(t)$; that is, $L(t)[T(t)f] = f$ for all $f \in L^2(\Omega)$. We note that $\{T(t)\}_{0\le t\le T}$ is a smooth family of bounded operators from H^{ℓ} to $H^{\ell+2} \cap D_L$, for $\ell \ge 0$ and that $\{L(t)\}_{0\le t\le T}$ is a smooth family of bounded operators from $H^{\ell+2} \cap D_L$ to H^{ℓ}, for $\ell \ge 0$. In fact, $L^{(j)}(t) = (\frac{d}{dt})^j L(t)$, $j \ge 0$, can be calculated by differentiating the coefficients of $L(t)$ with respect to time and if we let $T^{(j)}(t) = (\frac{d}{dt})^j T(t)$, we have that

$$T^{(1)}(t) = -T(t)\, L^{(1)}(t) T(t) .$$

We shall use the symbol C to denote a generic positive constant throughout this paper and we will define $\sum_{j=m_1}^{m_2} (\cdot) \equiv 0$ if $m_2 < m_1$.

II. Spacial Discretization Operators

We will assume that we have a finite dimensional subspace $S_h \subset L^2(\Omega)$ (associated with parameters $0 < h < 1$ and $r \geq 2$) and a sufficiently smooth family of selfadjoint, positive semidefinite operators $\{T_h(t)\}_{0\leq t\leq \tau}$ on $L^2(\Omega)$ that have range in S_h and are positive definite on S_h. We define $L_h(t)$ on S_h as the inverse of $T_h(t)\big|_{S_h}$ for each $0 \leq t \leq \tau$. Given $f \in L^2(\Omega)$, we will regard $T_h(t)f$ as an approximation to $T(t)f$. In fact we will require that the following be true:

$$\|(T^{(j)}(t) - T_h^{(j)}(t))f\| \leq C(j)h^{\ell+2}\|f\|_\ell , \quad j \geq 0 , \quad 0 \leq \ell \leq r-2 ; \tag{2.1}$$

here $T_h^{(j)}(t) = (\frac{d}{dt})^j T_h(t) : L^2(\Omega) \to L^2(\Omega)$. Finally, we will assume that there is a norm $\|\cdot\|_I$ on S_h that satisfies the following:

$$\|\varphi\|^2 \leq C\|\varphi\|_I^2 \leq C(L_h(t)\varphi,\varphi) \quad \text{for} \quad 0 \leq t \leq \tau , \ \varphi \in S_h , \tag{2.2}$$

$$|(L_h^{(j)}(t)\varphi_1,\varphi_2)| \leq C(j)\|\varphi_1\|_I\|\varphi_2\|_I \quad \text{for} \quad 0 \leq t \leq \tau, \varphi_1,\varphi_2 \in S_h \ \text{and} \ j \geq 0 ; \tag{2.3}$$

here $L_h^{(j)} = (\frac{d}{dt})^j L_h(t) : S_h \to S_h$.

We note that many of the well known Galerkin-type methods satisfy these assumptions. For a discussion, see Sammon [8].

We let $P_I(t) \equiv T_h(t)L(t) : D_L \to S_h$, $0 \leq t \leq \tau$, define a family of "elliptic projection" operators. We also let $P : L^2(\Omega) \to S_h$ denote the orthogonal $L^2(\Omega)$- projection onto S_h. (Note that $T_h = PT_hP$). If $w \in H^{\ell+2} \cap D_L$ for some $0 \leq \ell \leq r-2$, then

$$\|w - Pw\| \leq \|w - P_I w\| \leq Ch^{\ell+2}\|w\|_{\ell+2}, \quad 0 \leq t \leq \tau. \tag{2.4}$$

We let $P_I^{(j)}(t) = (\frac{d}{dt})^j P_I(t)$ for $j \geq 0$.

Suppose we choose $v \in H^r$ so that $u(t) \in H^r$ and $\|u(t)\|_r \leq C\|v\|_r$ for $0 \leq t \leq \tau$. Then setting $W(t) = P_I(t)u(t)$, we have the following:

$$\|u(t) - W(t)\| \leq C\, h^r\|v\|_r . \tag{2.5}$$

We also wish to see how well the time derivatives of W approximate those of u. To this end we study the following:

Proposition (2.1): If $w \in H^{\ell+2} \cap D_L$ for some $0 \le \ell \le r-2$, then

$$\|P_I^{(m)}(t)w\| \le C(m)h^{\ell+2}\|w\|_{\ell+2}, \quad 0 \le t \le \tau, \quad m > 0.$$

Proof: We see that (2.1) implies that

$$\|P_I^{(m)}(t)w\| \le \|\sum_{j=0}^{m} \binom{m}{j} T_h^{(j)} L^{(m-j)} w\|$$

$$\le \|\sum_{j=0}^{m} \binom{m}{j} T^{(j)} L^{(m-j)} w\| + C h^{\ell+2}\|w\|_{\ell+2}$$

$$= \|(\frac{d}{dt})^m TLw\| + C h^{\ell+2}\|w\|_{\ell+2} = C h^{\ell+2}\|w\|_{\ell+2}.$$

An easy application of the above result shows that if $0 \le \ell \le r-2$, $m \ge 0$ and $v \in H^{\ell+2+2m}$ is suitably chosen then

$$\|u^{(m)}(t) - W^{(m)}(t)\| \le C h^{\ell+2}\|v\|_{\ell+2+2m}, \quad 0 \le t \le \tau, \tag{2.6}$$

where $W^{(m)}(t) = (\frac{d}{dt})^m W(t)$. In particular, this shows that $W^{(m)}(t)$ is uniformly bounded in $L^2(\Omega)$ provided that u and v are suitably bounded.

At times, we will assume that the following condition holds:

B_h: $\|L_h^{(j)}(t)T_h(s)\|, \|T_h(s)L_h^{(j)}(t)\| \le C(j), \quad 0 \le s, t \le \tau, \quad j \ge 0.$

Estimates in Sammon [8] or Nassif and Descloux [7] show that this condition holds for various Galerkin-type methods if inverse assumptions are valid. The following result is a consequence of Condition B_h:

$$\|L_h(t)W^{(m)}(t)\| \le \sum_{j=0}^{m+1} \|u^{(j)}(t)\| \quad \text{for } m \ge 0, \ 0 < t \le \tau.$$

III. Time Discretizations

We now consider a method of computing approximations to the solution $u(t)$ of (1.1).

We begin by studying rational function approximations to the exponential e^{-x} on $\mathbb{R}^+$. It is well known that there are rational functions $r(x) = \frac{P(x)}{Q(x)}$ (P and Q are relatively prime polynomials) that satisfy the following conditions:

$$(3.1)\quad \begin{cases} \text{(i)} & Q(x) > 0 \quad \text{for} \quad x \geq 0\,, \quad Q(0) = 1\,, \\ \text{(ii)} & -1+\delta < Q^{-1}(x)P(x) < 1 \quad \text{for some} \quad \delta \geq 0\,, \quad x > 0\,, \\ \text{(iii)} & |r(x) - e^{-x}| \leq C\, x^{\nu+1} \quad \text{for some} \quad \nu \geq 1\,, \quad x \geq 0\,. \end{cases}$$

We will use P's and Q's that are no more than quadratics in our later work but for now, we assume that $P(x) = \sum_{\ell=0}^{\nu} p_\ell x^\ell$ and $Q(x) = \sum_{\ell=0}^{\nu} q_\ell x^\ell$ where $p_0 = q_0 = 1$. We have the following examples, where P and Q are of degree two or less:

(i) $p_1 = 0,\ q_1 = 1$ with $\delta = 1, \nu = 1$ (Backwards Euler)

(ii) $p_1 = -1/2,\ p_2 = 0,\ q_1 = 1/2,\ q_2 = 0$ with $\delta = 0, \nu = 2$ (Crank-Nicolson)

(iii) The family parametrized by $\lambda \geq 1/4,\ \lambda \neq 1/2$:

$p_1 = (2\lambda - 1),\ p_2 = (\lambda^2 - 2\lambda + 1/2),\ q_1 = 2\lambda,\ q_2 = \lambda^2$ with $\delta > 0,\ \nu = 2$

If $\lambda = \frac{1}{2}(1 + 1/\sqrt{3})$, $\nu = 3$ (Calahan)

(iv) $p_1 = p_2 = 0,\ q_1 = 1,\ q_2 = 1/2$ with $\delta = 1, \nu = 2$ (Padé)

(v) $p_1 = -1/3,\ p_2 = 0,\ q_1 = 2/3,\ q_2 = 1/6$ with $\delta > 0,\ \nu = 3$ (Padé)

(vi) $p_1 = -1/2,\ p_2 = 1/12,\ q_1 = 1/2,\ q_2 = 1/12$ with $\delta = 1,\ \nu = 4$ (Padé).

We are particularly interested in the cases where $\delta > 0$.

We will now show how property (3.1)(iii) can lead to a two point Taylor expansion used by Nassif and Descloux in [7].

Proposition (3.1): Suppose that $g(t)$ is a smooth function on $[0, t_0]$. Then for $0 \leq t \leq t_0$, we have that

$$(3.2)\quad \sum_{j=0}^{\nu} q_j (-t)^j g^{(j)}(t) = \sum_{j=0}^{\nu} p_j (-t)^j g^{(j)}(0) + \int_0^t K(t,s) g^{(\nu+1)}(s)\, ds$$

where $g^{(j)}(t) = (\frac{d}{dt})^j g(t)$ and

$$(3.3)\quad K(t,s) = \sum_{j=0}^{\nu} \frac{q_j}{(\nu-j)!} (-t)^j (t-s)^{\nu-j}\,.$$

Proof: From (3.1)(iii) we see that $Q(x)e^{-x} - P(x) = O(x^{\nu+1})$. Now take m derivatives, where $0 \le m \le \nu$, of each side of this equation and evaluate them at $x=0$. We see that

$$m!p_m = \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j}\, j!\, q_j \quad .$$

This gives (3.2) with $g(t) = t^m$, $0 \le m \le \nu$, and hence (3.2) if $g(t)$ is a polynomial of degree no more than ν.

We can expand a sufficiently smooth function $g(t)$ in a Taylor series of degree ν and apply our above work. This shows that (7) holds with the kernel given and completes the proof.

If we let $\{u_h(t)\}_{0\le t\le \tau} \subset S_h$ be defined by

$$u_{h,t} + L_h(t)u_h = 0 \ , \quad u_h(0) = v_h \ , \tag{3.4}$$

where $v_h \in S_h$ is some function "close" to v (for instance, $v_h = Pv$) then work by Sammon [8] shows that

$$\|u(t) - u_h(t)\| = O(h^r) \ , \quad 0 \le t \le \tau, \tag{3.5}$$

under certain conditions. Thus if we could approximate the solution $u_h(t)$ of (3.4) with a known small error, we could use (3.5) to show that our approximation is actually close to $u(t)$. We will use our two point Taylor polynomial to construct an approximation to u_h.

Let $0 < k < 1$ so that $Mk = \tau$ for some integer $M \ge 1$. We will study a method of approximating $u_h(k)$. Choose a rational function $r(x) = \frac{P(x)}{Q(x)}$ that satisfies (3.1)(i) through (3.1)(iii) and where the degrees of P and Q are two or less. (This implies that $\nu \le 4$.) If we note that

$$u_h^{(1)} = \frac{d}{dt} u_h = -L_h(t)u_h(t) \ ,$$

$$u_h^{(2)} = (\frac{d}{dt})^2 u_h = (L_h^2(t) - L_h^{(1)}(t))u_h(t),$$

then setting $g(t) = u_h(t)$, $0 \le t \le k$ and using (3.2) gives us the following:

$$\{I + q_1 kL_h(k) + q_2 k^2 (L_h^2(k) - L_h^{(1)}(k))\}u_h(k) \tag{3.6}$$

$$= \{I + p_1 kL_h(0) + p_2 k^2 (L_h^2(0) - L_h^{(1)}(0))\}u_h(0) + O(k^{\nu+1} u_h^{(\nu+1)}).$$

Thus if the quantity in the first set of braces is invertible, we might expect the following to be "close" to $u_h(k)$:

$$V = \{I + q_1 kL_h + q_2 k^2 (L_h^2 - L_h^{(1)})\}^{-1}(k)\{I + p_1 kL_h + p_2 k^2 (L_h^2 - L_h^{(1)})\}(0)\, v_h .$$

This scheme for approximating $u_h(k)$ will be the basis of a scheme for approximating $u_h(Nk)$ for any $N \geq 1$. It will be seen that the approximations can continue to be defined using operators that are constructed from P, Q and the family $\{L_h^{(j)}(t)\}_{j=0,1}$. We note that these schemes can be defined if the degrees of P and Q are higher than two, as was done in [7].

The solution $u_h(t)$ of (3.4) only plays a motivational role and will not enter in any way in the rest of this work. For purposes of our estimates, we will need some function in S_h that is uniformly close to $u(t)$ in the sense of (3.5) and $u_h(t)$ would be a possible candidate. We will use $W(t)$ for this purpose however, mainly because it is easy to estimate its time derivatives.

For $0 \leq n \leq M$, let $t_n = nk$, $L_n^{(j)} = L_h^{(j)}(t_n)(j \geq 0)$, $T_n^{(j)} = T_h^{(j)}(t_n)(j \geq 0)$, $P_n = P(-kL_n)$, $Q_n = Q(-kL_n)$, $\tilde{P}_n = P_n - p_2 k^2 L_n^{(1)}$, $\tilde{Q}_n = Q_n - q_2 k^2 L_n^{(1)}$. We now settle the question of $\tilde{Q}_n$'s invertibility on S_h.

Proposition (3.2): For k sufficiently small we have that

$$C_1(Q_n\varphi,\varphi) \leq (\tilde{Q}_n\varphi,\varphi) \leq C_2(Q_n\varphi,\varphi), \quad n \geq 0,\ \varphi \in S_h . \tag{3.7}$$

Proof: This follows immediately from the following inequality:

$$|(Q_n\varphi,\varphi) - (\tilde{Q}_n\varphi,\varphi)| = k^2|q_2(L_n^{(1)}\varphi,\varphi)| \leq Ck(Q_n\varphi,\varphi) .$$

We will assume that k is sufficiently small for the rest of this work. Then since Q_n is invertible due to $Q(x)$'s positivity on $\mathbb{R}^+$, the above Proposition shows that $\tilde{Q}_n$ is invertible.

We now return to our description of an approximation to the solution of (1.1). Given $v(x)$, we choose a $V^0 \in S_h$ that is close to v (for instance $V^0 = Pv$, although we will describe perhaps better possibilities later) and recursively define V^{n+1} $(0 \leq n < M)$ given V^n, by the following formula:

$$\tilde{Q}_{n+1} V^{n+1} = \tilde{P}_n V^n , \qquad 0 \leq n < M . \tag{3.8}$$

We expect V^n to be close to $u_h(t_n)$ which is in turn close to $u(t_n)$ (for $n \geq 0$) and we will derive corresponding estimates.

As noted before, this approximation scheme has been studied in an $L^2(\Omega)$ setting by Nassif and Descloux in [7]. We now see how computation of this scheme involves solving a new linear problem at each time step and why a more efficient variant would be desirable. We shall later study the variant suggested by Douglas, Dupont and Ewing in [4]. However, since the analysis of this variant requires estimates of the original scheme (the one defined by (3.8)) in new norms, we shall first present another analysis of this scheme. We shall also define a natural choice for V^0.

IV. Preliminary Error Estimates

We are primarily interested in how close V^n is to $W^n \equiv W(t_n)$ $(0 \le n \le M)$ since we already know (recall (2.6)) how close W^n is to $u^n \equiv u(t_n)$ $(0 \le n \le M)$. As noted before, these estimates are already known in the $L^2(\Omega)$-norm but we wish to study them in the (possibly) stronger norms given by $(Q_n(\cdot),(\cdot))^{1/2}$. This will allow us to study a variant of the scheme where the $(Q_n(\cdot),(\cdot))^{1/2}$ norm is in some sense a natural norm of the problem. We note that most of our work will go on in S_h in the next two sections.

Letting $E^n = V^n - W^n$ for $0 \le n \le M$, we see that

$$(4.1) \qquad Q_{n+1}E^{n+1} = P_{n+1}E^n + (P_n - P_{n+1})E^n + (\tilde{P}_n - P_n)E^n + (Q_{n+1} - \tilde{Q}_{n+1})E^{n+1} - (\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_nW^n) .$$

This will be an important error equation.

We note that (4.1) is of the form $QW = PV + F$ where $W,V,F \in S_h$ and Q and P are selfadjoint operators on S_h that satisfy the following:

$$(4.2) \quad \begin{cases} \text{(i)} & (Q\varphi,\varphi) > 0 , \quad 0 \ne \varphi \in S_h , \\ \text{(ii)} & ((Q-P)\varphi,\varphi) > 0 , \quad 0 \ne \varphi \in S_h , \\ \text{(iii)} & ((Q+P)\varphi,\varphi) > \delta(Q\varphi,\varphi), \quad 0 \ne \varphi \in S_h, \text{ for some } \delta \ge 0 . \end{cases}$$

(Of course (4.2)(i) through (4.2)(iii) follow from (3.1)(i) through (3.1)(iii).) This situation leads to the following:

Proposition (4.1): Let $QW = PV + F$ where Q and P are selfadjoint operators that satisfy (4.2). Then we have that

$$(4.3) \qquad (QW,W) \le (QV,V) - \delta((Q-P)V,V) + 2(F,W) .$$

Proof: We see that

$$(QW,W) = (PV,W) + (F,W) = (PV,Q^{-1}PV) + (Q^{-1}F,PV) + (F,W)$$
$$= (PV,Q^{-1}PV) + 2(F,W) - (Q^{-1}F,F)$$

and that $(PV,Q^{-1}PV) = (QV,V) - ((Q-P)(I+Q^{-1}P)V,V)$.

A simple calculation shows that $I+Q^{-1}P$ is selfadjoint in the inner product defined by $((Q-P)\cdot,\cdot)$ and hence has real eigenvalues. Thus if μ is an eigenvalue associated with the eigenfunction φ, (4.2)(iii) shows that $\mu(Q\varphi,\varphi) = (Q(I+Q^{-1}P)\varphi,\varphi) = ((Q+P)\varphi,\varphi) \ge \delta(Q\varphi,\varphi)$.

We conclude that the smallest eigenvalue, μ_m, satisfies $\mu_m \geq \delta$. Again using the self-adjointness, we see that

$$((Q-P)(I+Q^{-1}P)V,V) \geq \mu_m((Q-P)V,V) \geq \delta((Q-P)V,V)$$

which completes the proof.

We will want to apply Proposition (4.1) to (4.1) and obtain an estimate for $(Q_{n+1},E^{n+1},E^{n+1})$. In anticipation of this, we prove the following estimates.

Lemma (4.2): Suppose $n \geq 0$, $0 \leq m \leq C$ and we have Condition B_h if $Q(x)$ is quadratic. Then

$$\left.\begin{matrix} |((P_{n+m} - P_n)\varphi_1,\varphi_2)| \\ |((Q_{n+m} - Q_n)\varphi_1,\varphi_2)| \end{matrix}\right\} \leq C\,k(Q_n\varphi_1,\varphi_1)^{1/2}(Q_n\varphi_2,\varphi_2)^{1/2} . \tag{4.4}$$

Proof: We first note that (3.1)(ii) implies that the degree of $P(x)$ is no greater than the degree of $Q(x)$ and (3.1)(ii) and (3.1)(iii) imply that $Q(x)$ cannot be the constant 1. Also if we let $R(x) = 1+x$ if $q_2=0$ and $R(x) = 1+x+x^2$ if $q_2 \neq 0$, we see that $R^{-1}(x)Q(x)$, $R(x)Q^{-1}(x) \leq C$ for $x \geq 0$. Thus letting $R_j = R(kL_j)\,(0 \leq j \leq M)$, we have that

$$C_1(R_j\varphi,\varphi) \leq (Q_j\varphi,\varphi) \leq C_2(R_j\varphi,\varphi), \quad \varphi \in S_h, \quad j \geq 0 .$$

Thus it suffices to prove (4.4) with R_n-inner products on the right hand side.

We have the following estimates:

$$|((L_{n+m}-L_n)\varphi_1,\varphi_2)| \leq C\,k \sup_{|t_n-s|\leq mk} |(L_h^{(1)}(s)\varphi_1,\varphi_2)| \tag{4.5}$$

$$\leq C\,k\|\varphi_1\|_I\|\varphi_2\|_I \leq C\,k\,(L_n\varphi_1,\varphi_1)^{1/2}(L_n\varphi_2,\varphi_2)^{1/2} ,$$

$$|((L_{n+m}^2-L_n^2)\varphi_1,\varphi_2)| \leq |((L_{n+m}-L_n)\varphi_1,L_{n+m}\varphi_2)| + |(L_n\varphi_1,(L_{n+m}-L_n)\varphi_2)| \tag{4.6}$$

$$\leq C\,k\Big(\sup_{|t_n-s|\leq mk} \|L_h^{(1)}(s)T_n\|\Big)(1+\|L_{n+m}T_n\|)\|L_n\varphi_1\|\|L_n\varphi_2\| ;$$

note that Condition B_h was used in (4.6).

We now can use (4.5) and (4.6) to complete the proof.

Lemma (4.3): Suppose $n \geq 0$. Then

$$\left.\begin{matrix} |((\tilde{P}_n-P_n)\varphi_1,\varphi_2)| \\ |((\tilde{Q}_n-Q_n)\varphi_1,\varphi_2)| \end{matrix}\right\} \leq C\,k\,(Q_n\varphi_1,\varphi_1)^{1/2}(Q_n\varphi_2,\varphi_2)^{1/2} . \tag{4.7}$$

Proof: We note that there is nothing to prove unless $Q(x)$ is quadratic and that it suffices to prove (4.7) with R_n inner products on the right hand side, as was observed in the proof of Lemma (4.2). We have that

$$k^2 |(L_n^{(1)}\varphi_1,\varphi_2)| \le C k^2 \|\varphi_1\|_1 \|\varphi_2\|_1 \le C k^2 (L_n\varphi_1,\varphi_1)^{1/2}(L_n\varphi_2,\varphi_2)^{1/2} ,$$

which suffices to show (4.7).

We now study the truncation error term in (4.1) by comparing it to the true solution $u(t)$ of (1.1).

Proposition (4.4): Suppose that $v \in H^\mu$, $\mu = \max(2(\nu+1),r+2)$, is such that $\|u(t)\|_{r+2} \le C\|v\|_{r+2}$, $\|u^{(\nu+1)}(t)\| \le C\|v\|_{2(\nu+1)}$ for $0 \le t \le \tau$ and we have Condition B_h if $Q(x)$ is quadratic. Then for $0 \le n < M$ and $\varphi \in S_h$ we have that

$$|(\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_nW^n,\varphi)| \le C k (h^r\|v\|_{r+2} + k^\nu\|v\|_{2(\nu+1)}) (Q_n\varphi,\varphi)^{1/2} . \tag{4.8}$$

Proof: We note that $u_t = -Lu$, $u_{tt} = (L^2-L^{(1)})u = -L(I+T^{(1)})u_t$ for $0 \le t \le \tau$ and that Proposition (3.1) implies that

$$\|(u^{n+1}-q_1ku_t^{n+1}+q_2k^2u_{tt}^{n+1}) - (u^n-p_1ku_t^n+p_2k^2u_{tt}^n)\| \le C k^{\nu+1}\|v\|_{2(\nu+1)} . \tag{4.9}$$

We also have that $W^j = P_Iu^j = Pu^j + (P_I-P)u^j$,

$$L_jW = L_h(t_j)T_h(t_j)P\,L(t_j)\,u(t_j) = -Pu_t^j , \quad j \ge 0 , \tag{4.10}$$

$$\begin{aligned}(L_j^2-L_j^{(1)})W^j &= -L_j(P+T_j^{(1)})u_t^j \\ &= -L_j(P_Iu_t^j+P_IT^{(1)}(t_j)u_t^j) + L_j(P_I-P)u_t^j \\ &\quad + L_j(P_I-P)T^{(1)}(t_j)u_t^j + L_jP(T^{(1)}(t_j)-T_j^{(1)})u_t^j , \quad j \ge 0 ,\end{aligned} \tag{4.11}$$

and $-L_jP_I(u_t^j + T^{(1)}(t_j)u_t^j) = P\,u_{tt}^j$.

We now use these facts to see that

$$\begin{aligned}|(\tilde{Q}_{n+1}W^{n+1}-\tilde{P}_nW^n,\varphi)| &\le C k^{\nu+1}\|v\|_{2(\nu+1)}\|\varphi\| + C\Big(\int_{t_n}^{t_{n+1}} \|W^{(1)}(s)-Pu^{(1)}(s)\|ds\Big)\|\varphi\| \\ &+ C\{\|(P_I-P)u_t^{n+1}\| + \|(P_I-P)u_t^n\| + \|(P_I-P)T^{(1)}(t_{n+1})u_t^{n+1}\| + \|(P_I-P)T^{(1)}(t_n)u_t^n\| \\ &+ \|(T^{(1)}(t_{n+1})-T_{n+1}^{(1)})u_t^{n+1}\| + \|(T^{(1)}(t_n)-T_n^{(1)})u_t^n\|\}\, k^2(\|L_{n+1}\varphi\| + \|L_n\varphi\|) \\ &\le C k (h^r\|v\|_{r+2} + k^\nu\|v\|_{2(\nu+1)})\{\|\varphi\|^2 + k^2\|L_n\varphi\|^2\}^{1/2}\end{aligned}$$

which gives our result.

We can now put these results together and demonstrate a bound in a $Q^{1/2}$-norm for the difference between V^n and W^n.

Theorem (4.5): Suppose that v is sufficiently smooth and compatible (the hypotheses of Proposition (4.4) would suffice) and that we have Condition B_h if $Q(x)$ is quadratic. Then for $0 \le N \le M$,

$$\|Q_N^{1/2}(V^N - W^N)\| \le C(h^r + k^\nu)\|v\|_\mu + C\|Q_0^{1/2}(V^0 - P_I v)\| . \tag{4.12}$$

Proof: Let $0 \le n < M$. An application of the results of this section to the terms on the right of (4.1) shows that

$$(Q_{n+1}E^{n+1},E^{n+1}) \le (1+Ck)(Q_nE^n,E^n) + Ck(k^\nu + h^r)^2\|v\|_\mu^2 ,$$

which gives the result.

We will now examine possibilities for the starting function V^0. We require that $V^0 - P_I v$ be bounded by $C(k^\nu + h^r)$ in the $Q_0^{1/2}$-norm if we wish a comparable error in (4.12). We can always let $V^0 = P_I v$ but we note that the approximation scheme defined by (3.8) never requires that we determine T_h applied to any function. In applications, this would amount to a special, expensive calculation required only at the beginning. There is another choice for V^0 that involves solving a system with $\tilde{Q}_0$ (or even Q_0). Since such systems have to be solved anyway to take the first step, this would seem to be a better approach.

We have

Proposition (4.6): Let $v \in H^r \cap D_L$, $L(0)v \in D_L$ and define $V^{0,1}$ and $V^{0,2}$ in S_h by the following:

$$\begin{cases} \tilde{Q}_0 V^{0,1} = P(v + q_1 kL(0)v + q_2 k^2(L^2(0) - L^{(1)}(0))v) \\ Q_0 V^{0,2} = P(v + q_1 kL(0)v + q_2 k^2 L^2(0)v) = P\,Q(kL(0))v . \end{cases} \tag{4.13}$$

Then

$$\left.\begin{aligned} \|Q_0^{1/2}(V^{0,1} - P_I v)\| \\ \|Q_0^{1/2}(V^{0,2} - P_I v)\| \end{aligned}\right\} \le C h^r \|v\|_r . \tag{4.14}$$

Proof: We observe that

$$(L_0^2 P_I v - L^2(0)v,\varphi) = ((I - P_I)L(0)v, L_0\varphi)$$

$$(L_0^{(1)} P_I v - L^{(1)}(0)v,\varphi) = ((T^{(1)}(0) - T_0^{(1)})L(0)v, L_0\varphi) + ((P_I - I)T^{(1)}(0)L(0)v, L_0\varphi).$$

Thus, if we let $\varphi = v^{0,1} - P_I v$ then

$$\begin{aligned} \|Q_0^{1/2}(v^{0,1} - P_I v)\|^2 &\le C\|\tilde{Q}_0^{1/2}(v^{0,1} - P_I v)\|^2 \\ &\le |(v - P_I v, \varphi)| + q_2 k^2 C h^r \|v\|_{r+2} \|L_0 \varphi\| \\ &\le C h^r \|v\|_r \|Q_0^{1/2} \varphi\| . \end{aligned}$$

This completes the proof of the first part of (4.14) and the second part follows even more simply from the above observations.

This result completes our error analysis for the approximation scheme defined by (3.8) with v^0 defined by either equation of (4.13). We will call this the base scheme in the sequel.

We note that if $v^{0,2}$ is defined by

$$\bar{Q}(kL_0)v^{0,2} = P\bar{Q}(kL(0))v \tag{4.15}$$

where $\bar{Q}(x)$ is a polynomial that satisfies $0 < \sup_{0 \le x < \infty} \bar{Q}(x)/Q(x) < \infty$, an estimate like (4.14) will still hold. This modification might prove useful if $\bar{Q}(kL_0)$ is a preconditioning operator for the kind of linear system solving techniques we will study in the next section.

V. A Variant of the Base Scheme

As we noted before, the calculation of the base scheme involves the solution of a new linear problem at each time step. We wish to study a variant of this scheme where we only approximately solve the linear system at each time step. We propose to use an iterative technique for this purpose which, as we will see, can be provided with a good initial guess for the true solution.

If we are at a point in our calculations where we have several accurate approximations to the function $u(t)$ at previous time steps, it can be seen that there is an extrapolation of these values that yields just as good an approximation at the next time step. The smoothness of $u(t)$ makes this possible. This extrapolation could be used as an initial guess for an iterative procedure. But this observation raises a question. Since even the exact solution of the system which we are approximately solving is no closer to u (in the sense of order) than the extrapolated guess, why iterate at all? If we did no iterations and used this procedure as an algorithm to generate further approximations, errors would grow and the approximations would deteriorate. Such an algorithm is not stable. Of course, the base algorithm (solve the system exactly and forget about iterations) can easily be shown to be stable although we will not formally state this result. Also, it will not be too hard to see that if we make an error in approximately solving the system that is of the order of the local truncation error and that is in some sense independent of the initial guess, then the algorithm is stable and gives accurate approximations. For the iterative schemes that we will consider, this strategy requires a quantity of iterations that is on the order of the logarithm of the total number of time steps, per time step. However, there is a more efficient strategy available if the polynomials $P(x)$ and $Q(x)$ have the correct properties. If one does in fact give a good initial guess to the iterative scheme and then iterates only a certain number of times per time step (a number that is independent of the total number of steps), then even though accuracy is not improved, the resulting algorithm is stable and generates accurate approximations for u. This phenomenon was first observed in [4] in relation to the Crank-Nicolson scheme. We will give arguments in this section that show that these results hold for schemes that have the

right kind of dissipation; that is, P and Q are such that $\delta > 0$. Similar results can be proven for polynomial pairs that are just stable ($\delta = 0$) but the condition $k \leq C h^2$, for some constant C, is required. This condition introduces dissipation and was used in [4].

We begin by discussing the properties of a particular type of preconditioned iterative technique for solving linear systems. We will assume that we are working in a finite dimensional space H with an inner product $(\cdot,\cdot)_H$ and a norm $\|\cdot\|_H = (\cdot,\cdot)_H^{1/2}$. Suppose A is a positive definite selfadjoint operator on H and we wish to find an approximation to the vector x that satisfies $Ax = y$, where y is known. We will also assume that the situation is such that we have another positive definite selfadjoint operator A_o at our disposal for which $A_o^{-1}z, z \in H$ is easy to find and for which we know the following spectral estimate:

$$\lambda_0(A_o z,z)_H \leq (Az,z)_H \leq \lambda_1(A_o z,z)_H , \qquad z \in H , \tag{5.1}$$

where $0 < \lambda_0 \leq \lambda_1$ are known constants. Then there are methods which when given an initial guess $x^{(0)}$ for the solution x, generate a sequence of approximations $x^{(\alpha)}$, $\alpha \geq 1$, to x that have the following properties:

(i) The calculation of $x^{(\alpha+1)}$ given $\{x^{(j)}\}_{j=0}^{\alpha}$ only requires evaluating A and A_o, solving systems involving A_o and Hilbert space operations,

(ii) The sequence $x^{(\alpha)} \to x$ as $\alpha \to \infty$ geometrically in the following way. There is a decreasing function $0 \leq \gamma(\xi) < 1$ $(0 < \xi \leq 1)$ that satisfies $\gamma(1) = 0$ and which gives the rate of convergence of the iterative scheme in the $A_o^{1/2}$ norm:

$$\|A_o^{1/2}(x-x^{(\alpha)})\|_H \leq C_\lambda \gamma^\alpha\left(\frac{\lambda_0}{\lambda_1}\right)\|A_o^{1/2}(x-x^{(0)})\|_H , \qquad \alpha \geq 0 . \tag{5.2}$$

A given method may or may not actually use the spectrum estimation constants λ_0 and λ_1 in its calculations. We also note that (5.2) implies that the following estimate holds:

$$\|A_o^{1/2}(x-x^{(\alpha)})\|_H \leq C_\lambda\left(1-\gamma^\alpha\left(\frac{\lambda_0}{\lambda_1}\right)\right)^{-1}\gamma^\alpha\left(\frac{\lambda_0}{\lambda_1}\right)\|A_o^{1/2}(x^{(\alpha)}-x^{(0)})\|_H , \qquad \alpha > 0 . \tag{5.3}$$

The preconditioned conjugate gradient algorithm fits into this framework with $\gamma = (1-\xi^{1/2})/1+\xi^{1/2}$; see [1]. We also note that another example of such an algorithm is given by the following descent method. Let $\mu > 0$ and given $x^{(\alpha)}$ for some $\alpha \geq 0$, define $x^{(\alpha+1)}$ by the following:

$$A_o x^{(\alpha+1)} = \mu y + (A_o - \mu A)x^{(\alpha)} .$$

This method converges for certain values of μ and if we choose $\mu = \frac{2}{\lambda_1+\lambda_0}$ then we have a method that satisfies (i), (ii) above with $\gamma = (1-\xi)/1+\xi$.

We intend to use an iterative scheme with the properties described above to approximately solve the system (3.8) which defines our base scheme. The Hilbert space H will be S_h with the $L^2(\Omega)$-inner product and the above discussions outline possible error results. We will keep the conjugate gradient algorithm in mind since it offers a good convergence rate and it does not require the values of λ_0 and λ_1 in its calculations; they only enter into its error analysis.

We now must decide what to use as a preconditioning operator. We note that the contribution of the $L^{(1)}_{n+1}$ term in $\tilde{Q}_{n+1}$ is small, so we can ignore this term when constructing a preconditioner. We now discuss two possibilities:

A. Q_0 <u>as a preconditioning operator</u>.

This is a good choice if linear systems involving Q_0 are easy to solve. For instance, if $Q(x) = 1+q_1 x$ is linear then $Q_0 = I+q_1 kL_0$ has essentially the same structure as the L_0 operator and solving this type of problem is well studied. If $Q(x) = (1+\lambda x)^2$ as in the Calahan method, then solving systems with $Q_0 = (I+\lambda kL_0)^2$ only involves solving two successive systems with the $(I+\lambda kL_0)$ operator. Thus Q_0 is also a good preconditioner for this method. If $Q(x)$ is not a perfect square, the fourth order diagonal Padé scheme being a notable example, there are other methods for solving systems involving Q_0 that use complex arithmetic. Thus using Q_0 as a preconditioning operator is possible for these methods. We will offer an alternative in B however.

We observe that the following result contains (5.1) for Q_0 as a preconditioning operator.

<u>Proposition (5.1)</u>: Let $0 \le m,n \le M$ and assume Condition B_h if $Q(x)$ is quadratic. Then for $\varphi \in S_h$,

$$(5.4) \qquad (1+C|t_n-t_m+k|)^{-1}(Q_n\varphi,\varphi) \le (\tilde{Q}_m\varphi,\varphi) \le (1+C|t_n-t_m+k|)(Q_n\varphi,\varphi) .$$

Proof: We see that

$$(\tilde{Q}_m\varphi,\varphi) = (Q_n\varphi,\varphi) + ((Q_m-Q_n)\varphi,\varphi) + ((\tilde{Q}_m-Q_m)\varphi,\varphi)$$

and, using the techniques of Proposition (4.2) and (4.3), that

$$\left|((Q_m-Q_n)\varphi,\varphi) + ((\tilde{Q}_m-Q_m)\varphi,\varphi)\right| \leq C|t_n-t_m+k|(Q_n\varphi,\varphi) .$$

This gives the second inequality and the first is done in a similar fashion using Proposition (3.2).

Thus when we are using the iteration technique at the n-th time step, $n \geq 1$, we can always take $\lambda_0 = \lambda_1^{-1} = (1+C\,t_{n+1})^{-1}$ and we can expect an error reduction by a factor of at least $\gamma^{\alpha}((1+C\,t_{n+1})^{-2})$ after α iterations. Note that this implies that $\gamma \leq C\,t_{n+1}$ so that $\gamma \leq C\,k$ for the first few steps. Also if we are given an $\varepsilon > 0$, there is an $\alpha \geq 1$ (independent of n) so that $\gamma^{\alpha} \leq \varepsilon\, t_n^{1/2}$.

Now we consider another possibility for a preconditioning operator that is useful if $Q(x)$ is quadratic but not a perfect square. Let $\lambda > 0$ and set $S_n = I + \lambda k L_n$ for $0 \leq n \leq M$.

B. $S_0 = (I+\lambda k L_0)$ or S_0^2 as a preconditioning operator

We first note that it is easy to solve systems using these operators; that is, we only need to solve (perhaps successive) systems with the $(I+\lambda k L_0)$ operator. We can also prove the following result by the methods used in the proofs of Propositions (3.2), (4.2) and (4.3):

Proposition (5.2): Let $0 \leq m,n \leq M$ and ℓ be the degree of $Q(x)$. Suppose that Condition B_h is satisfied if $Q(x)$ is quadratic. Then for $\varphi \in S_h$, we have that

$$C_1(S_n^{\ell}\varphi,\varphi) \leq (\tilde{Q}_m\varphi,\varphi) \leq C_2(S_n^{\ell}\varphi,\varphi) . \tag{5.5}$$

Thus when we are using the iteration technique at the n-th time step, $n \geq 1$, we can always take $\lambda_0/\lambda_1 = C_1/C_2 < 1$ and obtain an error reduction by a factor of at least γ^{α} after α iterations. If we are given an $\varepsilon > 0$, there is an $\alpha \leq C \log(1/\varepsilon\, t_n^{1/2})$ so that $\gamma^{\alpha} \leq \varepsilon\, t_n^{1/2}$, where we assume that $\varepsilon\, t_n^{1/2} < 1$. We cannot assume that α is independent of n this time. However we note that if we do $\alpha_n \leq C|\log(t_n)| + C$ iterations at the n-th time step where n ranges between 1 and some N, $N>>1$, in the course of calculating an approximation to $u(t_N)$, then the average number of iterations per time step is

$$\frac{1}{N}\sum_{n=1}^{N}\alpha_n = \frac{M}{N}\left(\sum_{n=1}^{N}\alpha_n\frac{1}{M}\right) \le C .$$

Hence the <u>average</u> number of iterations per time step is independent of N, for large N.

We now gather these ideas. We will assume that we have chosen a preconditioning operator, which we will call ${}^{P}Q$ and we have Condition B_h if $Q(x)$ is quadratic. Thus we can assume that

(5.6) $$C_1({}^{P}Q\varphi,\varphi) \le (\tilde{Q}_n\varphi,\varphi) \le C_2({}^{P}Q\varphi,\varphi) \quad \text{for} \quad 0 \le n \le M, \quad \varphi \in S_h .$$

We also assume that we have an iterative linear system solving process $\mathbb{P}$ which uses this preconditioning operator. We now wish to use $\mathbb{P}$ to calculate approximations to the solution of $\tilde{Q}_n\chi = F$, assuming we have been given the right hand side F, an initial guess χ_o for χ and a tolerance $\beta_n > 0$. We will assume that there is an $\alpha_n = \alpha_n(\beta_n)$ so that if $\chi^{(\alpha_n)}$ is the α_n-th iterate of the process $\mathbb{P}$, then

(5.7) $$\|{}^{P}Q^{1/2}(\chi-\chi^{(\alpha_n)})\| \le \beta_n\|{}^{P}Q^{1/2}(\chi-\chi_o)\| .$$

Finally, we will make an assumption about the total number of iterations needed to achieve certain tolerances. If $\beta_n = \varepsilon\, t_n^{1/2}$ for $1 \le n \le N$, where $\varepsilon > 0$, then we will require that $\frac{1}{N}\sum_{n=1}^{N}\alpha_n \le C$ for large $N \le M$; that is, we only need finitely many iterations per time step, on the average, to achieve these tolerances.

With the process $\mathbb{P}$ at hand, we now formally state a variant of the algorithm stated in (3.8). First of all, given a $v(x)$, we choose a $U_0 \in S_h$ that is close to v and given a set $\{\beta_n\}_{n=1}^{M}$ of positive tolerances, we define U^{n+1} in terms of $\{U^j\}_{j=0}^{n}$, $0 \le n \le M-1$, in the following way. We use enough iterations of the process $\mathbb{P}$ (which uses the preconditioning operator ${}^{P}Q$) to generate an approximation U^{n+1} to the (true) solution $\bar{U}^{n+1}$ of the following system:

(5.8) $$\tilde{Q}_{n+1}\bar{U}^{n+1} = \tilde{P}_n U_n ,$$

where the error made is to be less than the tolerance β_{n+1}, in the sense of (5.7). We use

(5.9) $$Z_{n+1}(U) = \sum_{j=0}^{n}\gamma_{n+1,j}U^j$$

for certain coefficients $\{\gamma_{n+1,j}\}_{j=0}^{n}$, as an initial guess. (We will fix values for these coefficients later when it will be clearer what they need to be. Of course, letting $Z_{n+1}(U) = U^n$ is a possibility and in general, we will never use more than the past few values.)

If we redefine $E^n = U^n - W^n$ for $n \geq 0$, we note that we have the following important identity, an analogue of (4.1):

$$
\begin{aligned}
(5.10) \qquad Q_{n+1}E^{n+1} &= P_{n+1}E^n + (P_n - P_{n+1})E^n + (\tilde{P}^n - P^n)E^n \\
&\quad + (Q_{n+1} - \tilde{Q}_{n+1})E^{n+1} - (\tilde{Q}_{n+1}W^{n+1} - \tilde{P}_n W^n) \\
&\quad + \tilde{Q}_{n+1}(U^{n+1} - \tilde{U}^{n+1}), \qquad 0 \leq n \leq M-1 .
\end{aligned}
$$

We now analyze the error made by this kind of approximation algorithm. We will begin by studying a result that is easy to obtain but is not the best possible for our situation. We will briefly assume that we solve (5.8) to an error of $\beta_n = k^\nu$ for $1 \leq n \leq \nu-1$ (if $\nu \geq 2$) and to an error of $\beta_n = k$ for $\nu \leq n \leq M$. We note that these latter tolerances imply that for our types of processes $\mathbb{P}$, we must do on the order of $\log(M) = \log(\tau/k)$ iterations per time step in general. One might expect these tolerances to lead to good error estimates (if we use the appropriate initial guesses) and we will show that they indeed lead to good estimates. However we will later show that we can get the same type of estimates for a modified algorithm that only requires finitely many iterations per time step, on the average.

We will assume that $v(x)$ is sufficiently smooth and compatible for this discussion. In particular, this implies that we can take U^0 to be an approximation generated by $\mathbb{P}$ to either $V^{0,1}$ or $V^{0,2}$ (recall that these were defined in Proposition (4.6)) with an initial guess of zero and an error tolerance of $\beta_0 = h^r$. (However, recall (4.15).)

Our algorithm is of course not well defined and in fact will not obtain the accuracy claimed unless we make some special choices for the starting guesses required by the process $\mathbb{P}$. To be able to do this for the various schemes, we introduce some specific examples of the operators discussed in (5.9) as follows:

$$
\begin{aligned}
&Z^{(0)}_{n+1}(U) = 0 \quad \text{for} \quad 0 \leq n < M, \quad Z^{(1)}_{n+1}(U) = U^n \quad \text{for} \quad 0 \leq n < M, \\
&Z^{(2)}_{n+1}(U) = 2U^n - U^{n-1} \quad \text{for} \quad 1 \leq n < M, \\
&Z^{(3)}_{n+1}(U) = 3U^n - 3U^{n-1} + U^{n-2} \quad \text{for} \quad 2 \leq n < M, \\
&Z^{(4)}_{n+1}(U) = 4U^n - 6U^{n-1} + 4U^{n-2} - U^{n-3} \quad \text{for} \quad 3 \leq n < M, \\
&Z^{(5)}_{n+1}(U) = 5U^n - 10U^{n-1} + 10U^{n-2} - 5U^{n-3} + U^{n-4} \quad \text{for} \quad 4 \leq n < M .
\end{aligned}
$$

We can now state this not quite best possible algorithm in its entirety:

Algorithm (1): Use $\mathbb{P}$ with the preconditioning operator P_Q to

(1) generate an approximation U^0 to either $V^{0,1}$ or $V^{0,2}$ using zero as an initial guess and a tolerance $\beta_0 = \frac{1}{2} h^r$;

(2) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z^{(1)}_{n+1}(U) = U^n$ as an initial guess and a tolerance $\beta_{n+1} = k^\nu$ in the range $1 \le n+1 \le \nu-1$;

(3) generate an approximation U^{n+1} to the (true) solution of (5.8) using $Z^{(\nu)}_{n+1}(U)$ as an initial guess and a tolerance $\beta_{n+1} = k$, in the range $\nu \le n+1 \le M$.

Again we note that since we are using a tolerance of $\beta_{n+1} = k$ for $\nu \le n+1 \le M$, we need on the order of $\log(M) = \log(\tau/k)$ iterations per time step, in general.

We use the techniques of Section IV to study this process via Equation (5.10). We first observe the following:

$$(5.11)\qquad \begin{aligned} |(\tilde{Q}_{n+1}(U^{n+1}-\bar{U}^{n+1}),E^{n+1})| &\le C\|Q_{n+1}^{1/2}E^{n+1}\|\|P_Q^{1/2}(U^{n+1}-\bar{U}^{n+1})\| \\ &\le C\|Q_{n+1}^{1/2}E^{n+1}\|\beta_{n+1}\|Q_{n+1}^{1/2}(U^{n+1}-Z^{(i)}_{n+1}(U))\| \\ &\le C\beta_{n+1}\|Q_{n+1}^{1/2}E^{n+1}\|(\|Q_{n+1}^{1/2}(E^{n+1}-Z^{(i)}_{n+1}(E))\| \\ &\qquad + \|Q_{n+1}^{1/2}(W^{n+1}-Z^{(i)}_{n+1}(W))\|), \end{aligned}$$

$$(5.12)\qquad \|Q_{n+1}^{1/2}(E^{n+1}-Z^{(i)}_{n+1}(E))\| \le C\sum_{j=n-i+1}^{n+1}\|Q_j^{1/2}E^j\|,$$

where $i=1$ or ν depending on n . If we have Condition B_h and a suitable $v(x)$, then

$$(5.13)\qquad \begin{aligned} \|Q_{n+1}(W^{n+1}-Z^{(i)}_{n+1}(W))\| &\le C\|W^{n+1}-Z^{(i)}_{n+1}(W)\| + Ck\|L_{n+1}(W^{n+1}-Z^{(i)}_{n+1}(W))\| \\ &\le Ck^i\Big(\sup_{\substack{0\le s\le t_{n+1}\\ 0\le j\le i}}\|W^{(j)}(s)\| + \sup_{\substack{0\le s\le t_{n+1}\\ 0\le j\le i-1}}\|L_h(s)W^{(j)}(s)\|\Big) \\ &\le Ck^i\|v\|_{2(i+1)} \end{aligned}$$

where again $i=1$ or ν depending on n .

Thus, we can show the following:

Theorem (5.3): If we generate a sequence of approximations $\{U^n\}_{n=0}^M$ using Algorithm (1), where $v \in H^\mu$ ($\mu = \max(2(\nu+1),2(r+1))$) is sufficiently smooth and compatible and we assume Condition B_h, then for $N \geq 0$, we have that

$$\|Q_N^{1/2}(U^N - W^N)\| \leq C\,(h^r + k^\nu)\|v\|_\mu \,. \tag{5.14}$$

Proof: Let $n \geq 0$. We have via (5.10) and our estimates that

$$(Q_{n+1}E^{n+1},E^{n+1}) \leq (1+Ck)(Q_nE^n,E^n) + Ck\|v\|_\mu^2(h^r+k^\nu)^2 + Ck\sum_{j=\max(n-\nu+1,0)}^{n-1}(Q_jE^j,E^j) \,. \tag{5.15}$$

Also if $0 \leq n \leq \nu-1$ then $\|Q_n^{1/2}E^n\| \leq C(k^\nu+h^r)$. These inequalities and (5.15) give our result.

Thus we have optimal order errors for Algorithm (1).

As a preliminary to analyzing an algorithm that requires only finitely many iterations per time step on the average, we prove some results for the following situation.

Suppose that a set of approximations $\{U^j\}_{j=n-i+1}^n$ to the functions $\{W^j\}_{j=n-i+1}^n$ are given, where $1 \leq i \leq \nu+1$ and $n \geq i-1$. Use the process $\mathbb{P}$ (with the preconditioning operator ^2Q) to generate an approximation U^{n+1} to the (true) solution $\bar{U}^{n+1}$ of (5.8) using $Z_{n+1}^{(i)}(U)$ as an initial guess and a tolerance $\beta_{n+1} > 0$.

By the analysis done so far, we already know the following.

Proposition (5.4): Suppose we have Condition B_h and $v(x)$ is sufficiently smooth and compatible. Then for any $\varepsilon > 0$, we have that

$$\begin{aligned}(Q_{n+1}E^{n+1},E^{n+1}) \leq{}& (1+Ck)(Q_nE^n,E^n) - \frac{\delta}{2}((Q_n-P_n)E^n,E^n) \\ &+ Ck((k^\nu+h^r)^2 + \varepsilon\beta_{n+1}^2k^{2i-2})\|v\|_\mu^2 \\ &+ \varepsilon\frac{\beta_{n+1}^2}{k}\sum_{j=n-i+2}^{n+1}\|Q_n^{1/2}(E^j-E^{j-1})\|^2 \,.\end{aligned} \tag{5.16}$$

The above result motivates the investigation of the equation stated in the following result.

Proposition (5.5): Let $QW = PV+F$ where Q and P are selfadjoint operators on S_h. Then we have that

(5.17) $$((Q+P)(W-V),W-V) + ((Q-P)W,W) = ((Q-P)V,V) + 2(F,W-V) .$$

Suppose further that Q and P satisfy (4.2) (iii). Then

$$\delta(Q(W-V),W-V) + ((Q-P)W,W) \leq ((Q-P)V,V) + 2(F,W-V).$$

We now apply Proposition (5.5) to (5.10). If we can enforce a certain important condition, namely that $\delta > 0$ for our polynomials $P(x)$ and $Q(x)$ (recall (3.1)), we can obtain an estimate that will allow us to analyze the last term in (5.16).

Proposition (5.6): Suppose we have Condition B_h, $v \in H^{\mu}$ ($\mu = \max(2r+2,2\nu+2)$) is sufficiently smooth and compatible and $\delta > 0$. Then there are constants $\beta^* > 0$ and $C_\delta > 0$ so that if $\beta_{n+1} \leq \beta^*$, we have that

(5.18) $$C_\delta \|Q_{n+1}^{1/2}(E^{n+1}-E^n)\|^2 + ((Q_{n+1}-P_{n+1})E^{n+1},E^{n+1})$$
$$\leq (1+Ck)((Q_n-P_n)E^n,E^n) + Ck^2(Q_{n+1}E^{n+1},E^{n+1}) + Ck^2(Q_nE^n,E^n)$$
$$+ Ck^2\{(h^r+k^\nu)^2 + \beta_{n+1}^2 k^{2i-2}\}\|v\|^2 + C\beta_{n+1}^2 \sum_{j=n-i+2}^{n} \|Q_j^{1/2}(E^j-E^{j-1})\|^2 .$$

Proof: Our usual techniques, with different uses of the arithmetic-geometric mean inequality, yield (5.18) with $(1+Ck)$ replaced by 1 and with the following extra term on the right hand side:

(5.19) $$|(((Q_{n+1}-P_{n+1}) - (Q_n-P_n))E_n,E_n)| .$$

Note that $Q(x)-P(x) = x + O(x^2)$ by the accuracy condition (3.1)(iii) and $Q(x)-P(x) > 0$ for $x > 0$. If we redefine $R(x) = x + |q_2-p_2|x^2$ then $C_1R(x) \leq Q(x)-P(x) \leq C_2R(x)$ for $0 < x < \infty$. Thus, under Condition B_h, the techniques of Section IV show that

(5.20) $$|(((Q_{n+1}-P_{n+1}) - (Q_n-P_n))E^n,E^n)| \leq Ck^2(L_nE^n,E^n) + Ck^3|q_2-p_2|\|L_nE^n\|^2$$
$$\leq Ck(R(kL_n)E^n,E^n) \leq Ck((Q_n-P_n)E^n,E^n) .$$

We could now combine (5.16) and (5.18) with suitable choices for the parameter i. We would then essentially find that $\|Q_N^{1/2}E^N\|$ is bounded by terms that are $O(h^r+k^\nu)$, terms that measure the initial error in the $Q_0^{i/2}$-norm and terms that measure the initial error in the $k^{-1/2}(Q_0-P_0)^{1/2}$-norm. The projection we have chosen for the initial data

(as given in Proposition (4.6)) is defined so that it is computable by the $\mathbb{P}$ process and so it leads to an initial error that is good in the $Q_0^{1/2}$-norm. Unfortunately it does not necessarily lead to one that is good in the $k^{-1/2}(Q_0-P_0)^{1/2}$-norm. We could now proceed in two ways. One approach is to let U^0 be $P_I v$ or look for another special approximation which is good in all the required norms. But since the process $\mathbb{P}$ would probably not be useful in generating such an approximation (the spectrum of P_Q does not bear the correct relationship to the spectrum of $L_h(0)$), a special process would be needed to generate only U^0. Since we would prefer to avoid this situation, we are led to using (5.18) in some other way. After all, it was only the direct use of (5.18) that gave this apparent problem.

We have the following result which combines (5.16) and a variant of (5.18); the latter uses multiplication by the time variable to avoid potential problems at time zero.

Proposition (5.7): Suppose that Condition B_h holds, $v \in H^{\mu}$ $(\mu=\max(2r+2,2\nu+2))$ is sufficiently smooth and compatible and $\delta > 0$. Then for each $\varepsilon > 0$ there is a $\beta^{**} > 0$ so that if $\beta_{n+1} \le \min(\beta^{**}, t_{n+1}^{1/2})$, we have the following where $\hat{C}$ is independent of ε:

$$
\begin{aligned}
&(Q_{n+1}E^{n+1},E^{n+1}) + \hat{C}\frac{t_{n+1}}{k}\|Q_{n+1}^{1/2}(E^{n+1}-E^n)\|^2 \\
&\quad + C\Big[\frac{t_{n+1}}{k}((Q_{n+1}-P_{n+1})E^{n+1},E^{n+1}) - \frac{t_n}{k}((Q_n-P_n)E^n,E^n)\Big] \\
&\quad \le (1+C\,k)(Q^nE^n,E^n) - (\delta/3)((Q_n-P_n)E^n,E^n) \\
&\qquad + C\,k\,(h^{2r}+k^{2\nu} + \beta_{n+1}^2 k^{2i-2})\|v\|_{\mu}^2 \\
&\qquad + \varepsilon\frac{t_{n+1}}{k}\sum_{j=n-i+2}^{n}\|Q_j^{1/2}(E^j-E^{j-1})\|^2 .
\end{aligned}
\tag{5.21}
$$

Proof: It is a rather straightforward computation using (5.16) and (5.18) to obtain (5.21) with $\delta/3$ replaced by $\delta/2$ and a term

$$C\varepsilon_1((Q_n-P_n)E^n,E^n)$$

on the right hand side where $\varepsilon_1 > 0$ is arbitrary. This gives the result.

We can now define and state results for our final algorithm:

Algorithm (2): Use $\mathbb{P}$ with the preconditioning operator P_Q to

(1) generate an approximation U^0 to either $V^{0,1}$ or $V^{0,2}$ using zero as an initial guess and a tolerance $\beta_0 = \frac{1}{2}h^r$.

(2) generate an approximation U^{n+1} to the (true) solution of (5.8) using $\mathcal{I}_{n+1}^{(n+1)}(U)$ as an initial guess and a tolerance $\beta_{n+1} \le \min(k^{\nu-n}, \tilde{\beta})$ in the range $1 \le n+1 \le \nu$, where $\tilde{\beta} > 0$ is small.

(3) generate an approximation U^{n+1} to the (true) solution of (5.8) using $\mathcal{I}_{n+1}^{(\nu+1)}(U)$ as an initial guess and a tolerance $\beta_{n+1} \le \min(\tilde{\beta}, t_{n+1}^{1/2})$ in the range $\nu+1 \le n+1 \le M$, where $\tilde{\beta} > 0$ is small.

We note that the important difference between this algorithm and Algorithm (1) is that by our assumptions on the process $\mathbb{P}$, we only have to iterate a fixed number of times at each time step, on the average. Thus we are demanding fewer iterations, but as we will soon see, we get the same convergence rates.

Theorem (5.8): Suppose that Condition B_h holds, $v \in H^{\mu}$ $(\mu=\max(r+2,2\nu+2))$ is sufficiently smooth and compatible and $\delta > 0$. Then there is a (computable) $\tilde{\beta} > 0$ so that the approximations $\{U^n\}_{n=0}^M$ generated by Algorithm (2) satisfy

$$\|Q_N^{1/2}(W^N-U^N)\| \le C(k^{\nu}+h^r)\|v\|_{\mu}, \quad N \ge 0,$$

$$\|W^N-U^N\|_I \le C\, t_N^{-1/2}(k^{\nu}+h^r)\|v\|_{\mu}, \quad N > 0,$$

$$\|u(t_N)-U^N\| \le C(k^{\nu}+h^r)\|v\|_{\mu}, \quad N \ge 0.$$

Note that we get a superconvergence result for W^N-U^N in the $\|\cdot\|_I$-norm for times bounded away from zero.

Proof: The proof is almost immediate from (5.21). We first see that

$$(Q_{n+1}E^{n+1},E^{n+1}) + \|Q_{n+1}^{1/2}(E^{n+1}-E^n)\|^2 + C((Q_{n+1}-P_{n+1})E^{n+1},E^{n+1}) \le C(h^{2r}+k^{2\nu})\|v\|_{\mu}^2 \tag{5.22}$$

for $1 \le n+1 \le \nu$, where we note the important role of the t variable in the case $n=0$.

If $N \ge \nu+1$, we set $i = \nu+1$ in (5.21), multiply the inequality by $e^{-C_1 t_{n+1}}$ for a suitable C_1 and sum over the range $\nu \le n \le N-1$. Then by choosing $\varepsilon > 0$ sufficiently small and using (5.22) for $0 \le n \le \nu$, we obtain the first result. The third result now easily follows.

The remaining result is obtained by noting that if $\varphi \in S_h$, $0 \le n \le M$ and $R(x) = x + |q_2-p_2|x^2$ then

$$\|\varphi\|_I \le C(I_n\varphi,\varphi) \le \frac{C}{k}(R(kL_n)\varphi,\varphi) \le \frac{C}{k}((Q_n-P_n)\varphi,\varphi).$$

Thus Algorithm (2) generates accurate approximations to the solution $u(t)$ of (1.1) if the polynomial pair is such that $\delta > 0$, implying that dissipation effects are present. As was noted before, similar results can be proven for stable polynomial pairs (for which $\delta = 0$) if dissipation is introduced by requiring that $k \leq C h^2$.

VI Computational Considerations

We conclude this paper with a few remarks concerning the computational aspects of Algorithm (2) for quadratic $P(x)$ and $Q(x)$. Suppose that the functions $\{\varphi_i\}_{i=1}^J$ form a basis for S_h. Moreover, suppose that these functions have been chosen so that linear systems on $\mathbb{R}^J$ involving the matrices

$$A_m = [(L_m\varphi_i,\varphi_j)]_{i,j=1}^J \ , \quad A_m^{(1)} = [(L_m^{(1)}\varphi_i,\varphi_j)]_{i,j=1}^J \ , \quad 0 \le m \le M \ ,$$
$$G = [(\varphi_i,\varphi_j)]_{i,j=1}^J$$

have acceptable computational properties. We now wish to examine Algorithm (2) in this context.

We begin this discussion by making some definitions and identifications. Let $0 \le m \le M$. If $\varphi = \sum \zeta_i\varphi_i \in S_h$ then

$$G\underline{\zeta} = [(\varphi,\varphi_j)]_{j=1}^J \ , \qquad L_m\varphi = \sum_{j=1}^J [G^{-1}A_m\underline{\zeta}]_j\varphi_j \ . \tag{6.1}$$

Thus if $\bar{U} = \sum \zeta_j\varphi_j$ satisfies $\tilde{Q}_m\bar{U} = \varphi$ then

$$(I + q_1kG^{-1}A_m + q_2k^2(G^{-1}A_mG^{-1}A_m - G^{-1}A_m^{(1)}))\underline{\zeta} = \underline{\xi} \tag{6.2}$$

or, equivalently,

$$B_m\underline{\zeta} \equiv (G + q_1kA_m + q_2k^2A_mG^{-1}A_m - q_2k^2A_m^{(1)})\underline{\zeta}$$
$$= G\underline{\xi} = [(\varphi,\varphi_j)]_{j=1}^J \ . \tag{6.3}$$

We note that (6.3) involves a symmetric, positive definite matrix B_m. We now let P_Q be one of the polynomials in kL_0 discussed in Section V. Let $P_Q(x) = 1 + \tilde{q}_1x + \tilde{q}_2x^2$ define its coefficients and let

$$P_B = G + \tilde{q}_1kA_0 + \tilde{q}_2k^2A_0G^{-1}A_0 \ ,$$

which is also a symmetric, positive definite matrix on $\mathbb{R}^J$. Finally, rewriting (5.6) gives us the following:

$$C_1(P_B\underline{\eta},\underline{\eta}) \le (B_m\underline{\eta},\underline{\eta}) \le C_2(P_B\underline{\eta},\underline{\eta}) \ , \quad \underline{\eta} \in \mathbb{R}^J \tag{6.4}$$

where $(\cdot,\cdot)$ is the usual inner product on $\mathbb{R}^J$.

Thus we see that S_h with the $L^2(\Omega)$-inner product and $\mathbb{R}^J$ with the $(G\cdot,\cdot)$ inner product are unitarily equivalent under the identification of φ_i with the i-th canonical basis function in $\mathbb{R}^J$. This implies that any process $\mathbb{P}$ on S_h that solves systems

involving Q_m using P_Q as a preconditioner is formally equivalent to a process $\mathbb{P}^J$ on $\mathbb{R}^J$ that solves systems involving $G^{-1}B_m$ using $G^{-1}P_B$ as a preconditioner. Since a process will usually be given on $\mathbb{R}^J$ in practice, this identification defines the corresponding process $\mathbb{P}$ on S_h.

How do we compute the coefficients $\{\zeta_i^n\}_{1\le i\le J, 0\le n\le M}$ for the basis expansion of the functions $\{U^n\}_{0\le n\le M} = \{\sum \zeta_i^n \varphi_i\}$ defined by Algorithm (2)? The linear system used to compute U^0 has $\varphi = Pf$ as a right hand side where $f \in L^2(\Omega)$. If $\varphi = \sum \xi_i^f \varphi_i$ then

$$G\underline{\xi}^f = [(Pf,\varphi_j)]_{j=1}^J = [(f,\varphi_j)]_{j=1}^J \quad . \tag{6.5}$$

Thus $G\underline{\xi}^f$ can be calculated by taking inner products with f and $\underline{\xi}^f$ can be found by solving a system involving G. The linear system used to compute U^{n+1} for some $n \ge 0$ has $\tilde{P}_n U^n = \sum \xi_i^n \varphi_i$ as a right hand side. If we know $\underline{\zeta}^n$ (the coefficients for U^n) then

$$G\underline{\xi}^n = (G + P_1 kA_n + P_2 k^2 (A_n G^{-1} A_n - A_n^{(1)}))\underline{\zeta}^n \, , \tag{6.6}$$

so that $G\underline{\xi}^n$ can be calculated by solving one system involving G and $\underline{\xi}^n$ can be calculated by solving two. Thus we can compute the right hand sides of the various systems. If we know $\{\underline{\zeta}^j\}_{j=0}^n$, the initial guess for the iteration procedure is easy to calculate. Then to find $\underline{\zeta}^{n+1}$ via the process $\mathbb{P}^J$, we may have to evaluate $G^{-1}B_{n+1}$ and $G^{-1}P_B$, solve systems involving $G^{-1}P_B$ and do various Hilbert space operations in $\mathbb{R}^J$ with the $(G\cdot,\cdot)$ inner product. Evaluating $G^{-1}B_{n+1}$ or $G^{-1}P_B$ is straightforward. (However, note that calculating

$$\langle G(G^{-1}B_{n+1})\underline{\eta}_1, \underline{\eta}_2\rangle = \langle B_{n+1}\underline{\eta}_1, \underline{\eta}_2\rangle \quad \text{for } \underline{\eta}_1, \underline{\eta}_2 \in \mathbb{R}^J$$

only means solving one system involving G). If we have to solve the system $G^{-1}P_B\underline{\eta} = \underline{\xi}$ in $\mathbb{R}^J$, we observe that this is equivalent to solving $P_B\underline{\eta} = G\underline{\xi}$. (This becomes convenient in many situations as $G\underline{\xi}$ may be easier to obtain than $\underline{\xi}$. This was the case in (6.5) and (6.6).) We must now solve systems that involve P_B. If $P_Q(x) = (1+\lambda x)^2$ for some $\lambda > 0$ for instance, then

$$P_B = (G+\lambda kA_0)G^{-1}(G+\lambda kA_0) \quad .$$

Thus solving $P_B\underline{\eta} = G\underline{\xi}$ in this case means solving two systems that involve the same matrix $(G+\lambda kA_0)$ and evaluating G once. If $P_Q(x)$ is not a perfect square, other techniques could be used to efficiently solve systems involving P_B.

The particular choice of the iterative process will determine which of the above considerations is relevant in the implementation of Algorithm (2).

References

[1] Axelsson, O., On preconditioning and convergence acceleration in sparse matrix problems, CERN European Organization for Nuclear Research, Geneva, 1974.

[2] Baker, G. A., Bramble, J. H. and Thomée, V., Single step Galerkin approximations for parabolic problems, Math. Comp. 31, (1977), pp. 818-847.

[3] Douglas, J., Jr. and Dupont, T., Alternating direction methods on rectangles, Numerical Solution of Partial Differential Equations - II, (B. Hubbard, ed.), Academic Press, New York, 1971.

[4] Douglas, J., Jr., Dupont, T. and Ewing, R., Incomplete iteration for time-stepping a Galerkin method for a quasilinear parabolic problem, SIAM J. Num. Anal., (to appear).

[5] Friedman, A., Partial Differential Equations, Robert E. Krieger Publishing Company, Huntington, New York, 1976.

[6] Lions, J. L. and Magenes, E., Nonhomogeneous Boundary Value Problems and Applications, II, Springer-Verlag, New York, 1973.

[7] Nassif, N. and Descloux, J., Stability study for time-dependent linear parabolic equations and its application to Hermitian methods, Topics in Numerical Analysis III, (J. Miller, Ed.), Academic Press, New York, 1977.

[8] Sammon, P., Approximations for parabolic equations with time dependent coefficients, Ph.D. Thesis, Cornell University, 1978.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 1958
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): (16) EFFICIENT HIGHER ORDER SINGLE STEP METHODS FOR PARABOLIC PROBLEMS, PART I
5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): (10) James H. Bramble and Peter H. Sammon
8. CONTRACT OR GRANT NUMBER(s): MCS78-09525; (15) DAAG29-75-C-0024; NSF MCS76-07236

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 7 - Numerical Analysis

11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below
12. REPORT DATE: (11) May 1979
13. NUMBER OF PAGES: 29

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (12) 33 pp
15. SECURITY CLASS. (of this report): UNCLASSIFIED
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(14) MRC-TSR-1958-PT-1

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(9) Technical summary Rept.,

18. SUPPLEMENTARY NOTES

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D. C. 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Parabolic Equations
Galerkin Methods
Higher Order Methods

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Some efficient, high order methods are discussed for approximating the solution of an initial boundary value problem for a homogeneous parabolic equation with time dependent coefficients. The methods are based on Galerkin-type approximations in the spacial variables and single step methods in the time variable. The equations defining the time stepping procedure are solved only approximately however. A preconditioned iterative technique is used for this purpose. The resulting algorithm is shown to produce optimal order approximations using only the order of work required by the single step method applied to the parabolic problem with time independent coefficients.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

R221 200